Machine Learning 1

Lecture 4.2 - Supervised Learning
Bias Variance Decomposition

Erik Bekkers

(Bishop 1.5.5, 3.2)

Slide credits: Rianne van den Berg

Image credit: Kirillm | Getty Images

Expected Loss for Regression

Frequentist viewpoint of model complexity

for a given $(\underline{x}, t) \sim p(\underline{x}, t)$

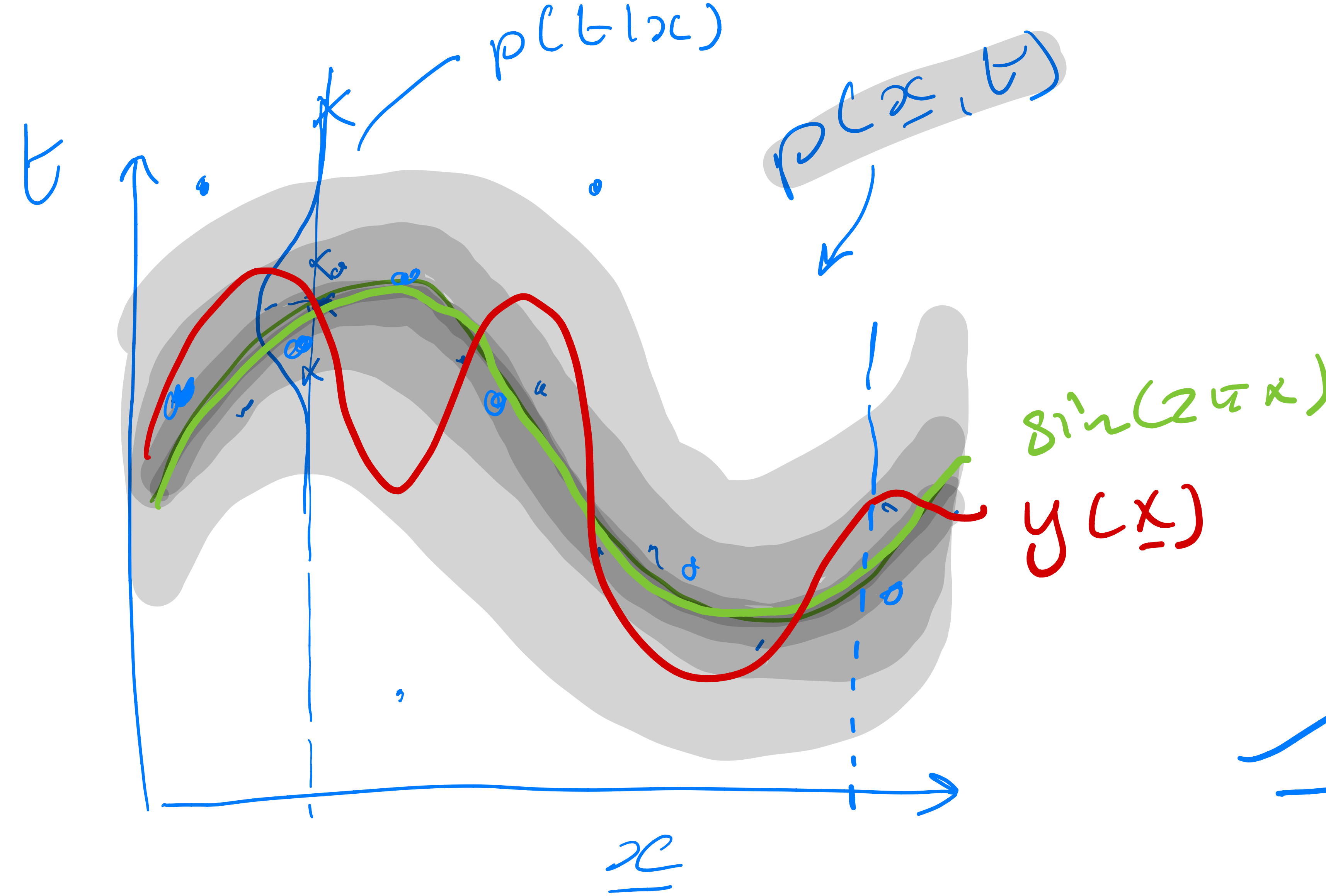
- Regression loss function: $L(t, y(\mathbf{x})) = (t - y(\underline{x}))^2$

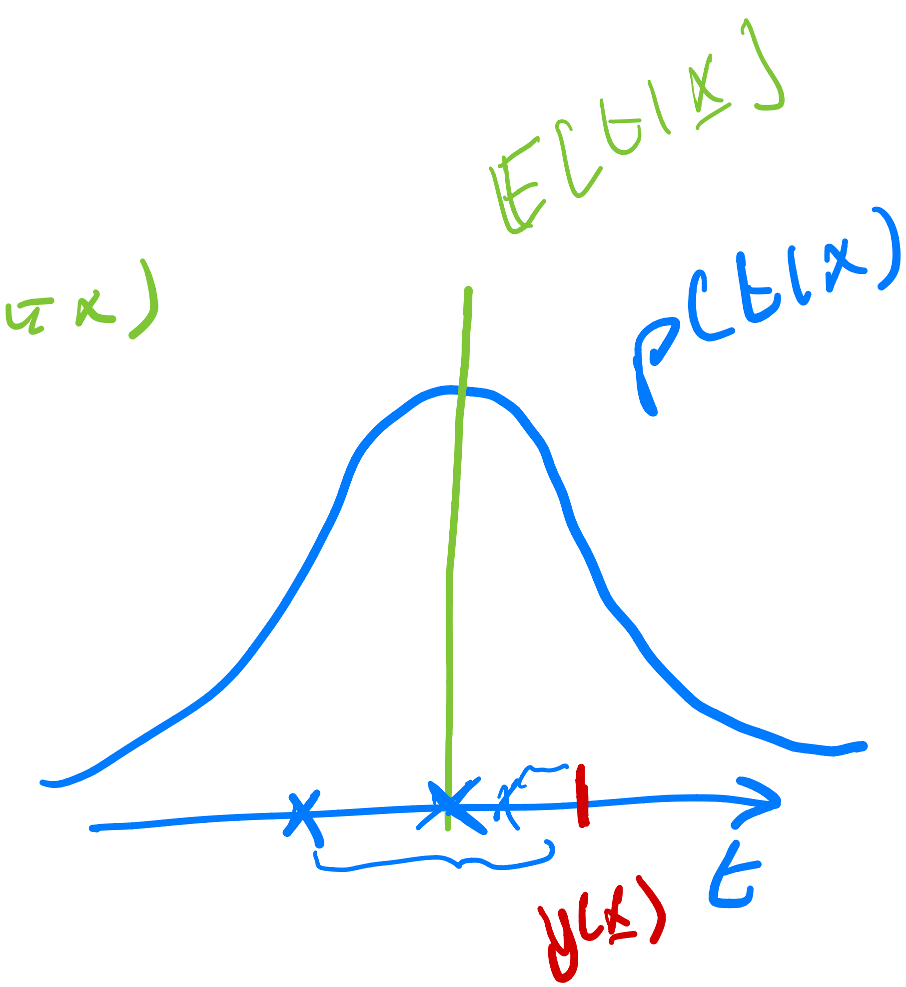
- Expected loss: $\mathbb{E}[L(t, y(\mathbf{x}))] = \iint (t - y(\underline{x}))^2 \, p(\underline{x}, t) \, d\underline{x} \, dt$

- Optimal y(**x**) minimizes $\mathbb{E}[L(t, y(\mathbf{x}))]$

$$y(\underline{x}) = \mathbb{E}[t \mid \underline{x}]$$

$t = \sin(2\pi x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \beta^{-1})$





For fixed $\underline{x}$, the expected loss

$$\mathbb{E}[L(t, y(\underline{x}))] = \int (t - y(\underline{x}))^2 \, p(t|x) \, dt$$

minimize w.r.t. $y(x)$

$\frac{d}{dy(x)} \ldots = 0$

$$\Longrightarrow \int y(\underline{x}) \, p(t|\underline{x}) \, dt = \int t \, p(t|\underline{x}) \, dt$$

regression function $\boxed{y(\underline{x}) = \mathbb{E}[t|\underline{x}]}$

Expected Loss for Regression

- Decomposition of expected loss:

verify Bishop 1.5.5

$$\mathbb{E}[L] = \int\int (\underbrace{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]} + \underbrace{\mathbb{E}[t|\mathbf{x}] - t})^2 \, p(\mathbf{x}, t)\mathrm{d}t\mathrm{d}\mathbf{x}$$

$$= \int \underbrace{(y(\mathbf{x}) - E[t|\mathbf{x}])^2}_{\text{bias}} p(\mathbf{x})\mathrm{d}\mathbf{x} + \underbrace{\int \mathrm{var}[t|\mathbf{x}]p(\mathbf{x})\mathrm{d}\mathbf{x}}_{\text{due to intrinsic noise}}$$

Minimizing the Expected Loss

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}) \mathrm{d}\mathbf{x} + \int \mathrm{var}[t|\mathbf{x}] p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

- Optimal solution is $y(\mathbf{x}) =$ $\mathbb{E}[t|x]$ (unknown)
- Only finite dataset observed: $\{(x_1, t_1), \ldots, (x_N, t_N)\} = D_1$
  $= D_2$
  $\vdots$
- Frequentist approach: estimate $y_D(\mathbf{x}) = y(\mathbf{x}, \mathbf{w}^*)$ based on dataset D
- Estimate performance of learning algorithm by averaging the expected loss over learned $y_D(\mathbf{x})$ for different datasets D

$$\mathbb{E}_D[\left(y_D(\mathbf{x}) - \mathbb{E}[t\,|\,\mathbf{x}]\right)^2]$$

Bias-Variance Decomposition

$$\mathbb{E}[\mathbb{E}_D[L]] = \int E_D[\left(y_D(\mathbf{x}) - \mathbb{E}[t\,|\,\mathbf{x}]\right)^2]p(\mathbf{x})\mathrm{d}\mathbf{x} + \int \mathrm{var}[t\,|\,\mathbf{x}]p(\mathbf{x})\mathrm{d}\mathbf{x}$$

- Bias-Variance decomposition:

$$\mathbb{E}_D[\{y_D(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2] = \mathbb{E}_D[\{\underline{y_D(\mathbf{x}) - \mathbb{E}_D[y_D(\mathbf{x})]} + \underline{\mathbb{E}_D[y_D(\mathbf{x})] - \mathbb{E}[t|\mathbf{x}]}\}^2] =$$

$$= \mathbb{E}_D\left[(y_D(x) - \mathbb{E}_D[y_D(x)])^2 + (\mathbb{E}_D[y_D(x)] - \mathbb{E}[t|x])^2 + \cancel{\{y_D(x) - \mathbb{E}_D[y_D(x)]\}\{\mathbb{E}_D[y_D(x)] - \mathbb{E}[t|x]\}}\right]$$

- Expected loss decomposition: $\mathbb{E}[\mathbb{E}_D[L]] = (\mathrm{bias})^2 + \mathrm{variance} + \mathrm{noise}$

$$(\text{bias})^2 = \int (\mathbb{E}_D[y_D(x)] - \mathbb{E}[t|x])^2 p(x)\,dx$$

$$\text{variance} = \int \mathbb{E}_D\left[(y_D(x) - \mathbb{E}_D[y_D(x)])^2\right] p(x)\,dx$$

$$\text{noise} = \int \mathrm{var}[t|\mathbf{x}]p(\mathbf{x})\mathrm{d}\mathbf{x}$$

Bias-Variance Decomposition: Example

- Generate L datasets of N points: = 100 = 25

  $x \sim U(0,1)$

  $t = \sin(2\pi x) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \alpha^{-1})$

  $\mathbb{E}[t|x] =$ sin(2πx)

- L predictions with 24 Gaussian basis functions

  $y^{(l)}(x) = (\mathbf{w}^{(l)})^T \boldsymbol{\phi}(x)$

  $$E_D = \frac{1}{2}\sum_{i=1}^{N}\{t_n - \mathbf{w}^T\boldsymbol{\phi}(x)\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

  $\mathbb{E}_D[y_D(x)] = \bar{y}(x)$

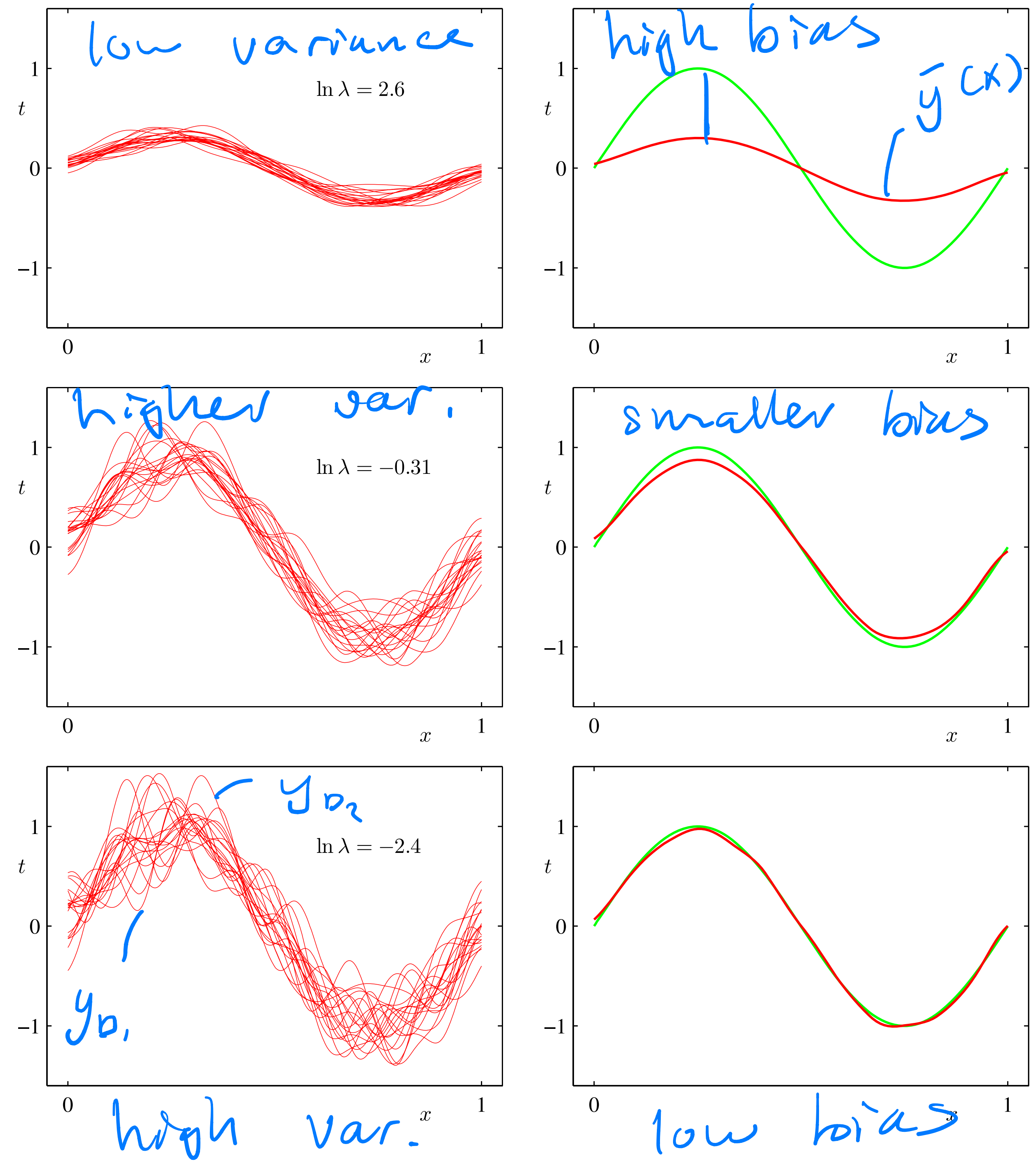


Figure: bias-variance decomposition (Bishop 3.5)

Bias-Variance Decomposition: Example

Estimate the bias and variance:

$\{x_1, x_2, \ldots, x_N\}$

- $$(\text{bias})^2 = \int \{\underbrace{\mathbb{E}_D[y_D(x)]}_{\bar{y}(x)} - \underbrace{\mathbb{E}[t|x]}_{\sin(2\pi x)}\}^2 p(x)\mathrm{d}x$$

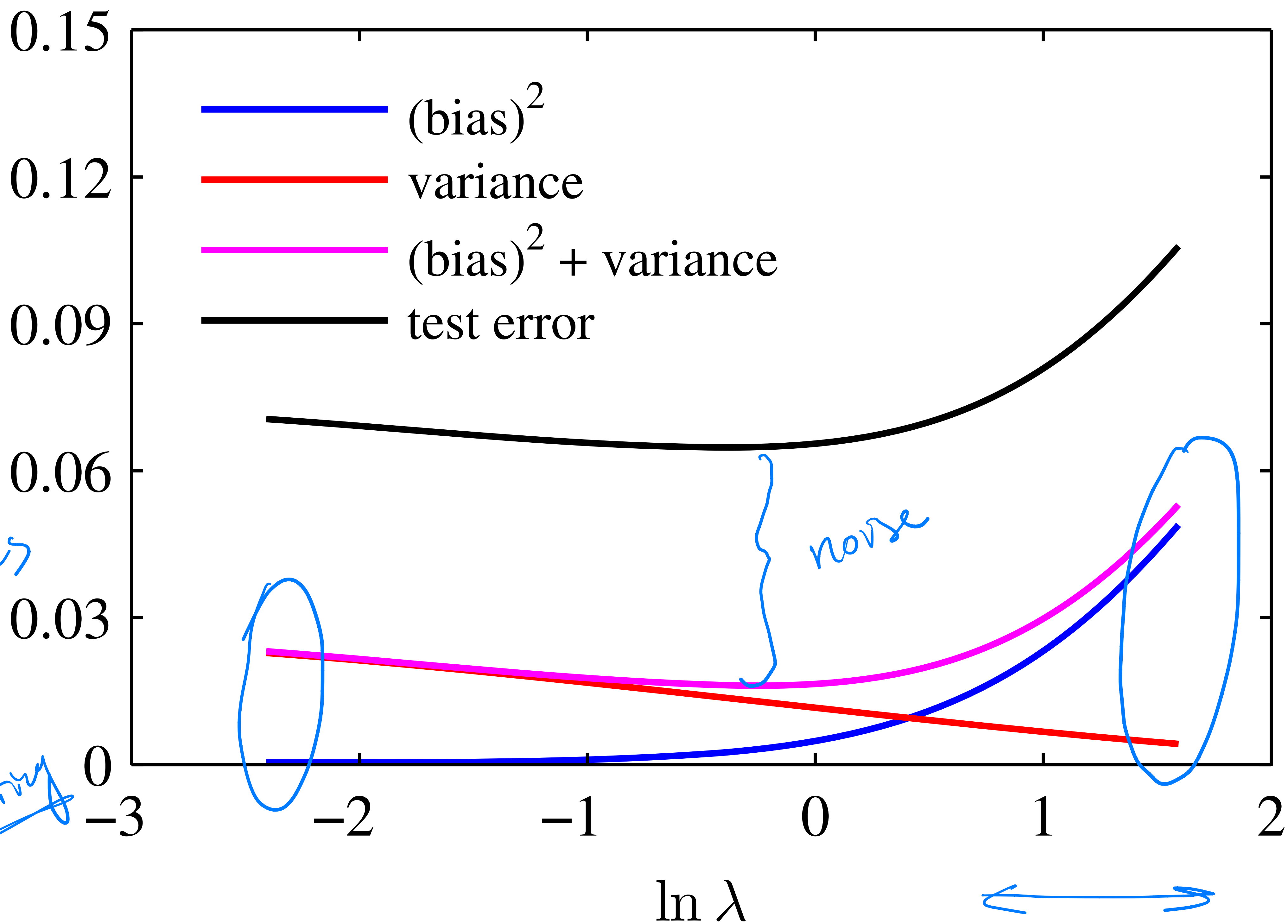
$$= \frac{1}{N}\sum_{n=1}^{N}\left(\bar{y}(x_n) - \mathbb{E}[t|x_n]\right)^2$$

$$\mathbb{E}_D[y_D(x)] = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x) = \bar{y}(x)$$

- $$\text{variance} = \int \mathbb{E}_D[\{y_D(x) - \mathbb{E}_D[y_D(x)]\}^2]p(x)\mathrm{d}x$$

$$= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left(y^{(l)}(x_n) - \bar{y}(x_n)\right)^2$$

Bias-Variance Decomposition: Example



low bias
high var
overfitting

underfitting
low var
high bias

Figure: bias-variance decomposition (Bishop 3.6) trade off
var / bias

Bias-Variance decomposition

- In practice we don't want to split our dataset into L datasets to determine the best model complexity (best value of λ)
- Better to keep large dataset,
  - Less overfitting.
  - Different optimal model complexity!
- Bayesian regression!